Bootstrapping in Stata

Bruce E. Hansen

Econ 706
Introduction: Uses of Bootstrap in Econometrics

- Standard Errors
  - Coefficient estimate
  - Function of estimates

- Confidence Intervals
  - Normal-based
  - Percentile
  - Bias-Corrected (BC)
  - Accelerated and Bias-Corrected (BC<sub>a</sub>)
  - Percentile-t

- Joint Tests

- Bootstrap for Quantile Regression

- Number of bootstrap replications
Example: Probit Model for Marriage

Sample: March 2009 CPS
Population: U.S. Black women in Midwest (n=433)
Percent Married: 37%
Probit for \textit{married} as a function of


\texttt{.probit mar age age2 education if bf, r}

- This calculates (robust) asymptotic standard errors
Probit regression

| mar  | Coef.    | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|------|----------|-----------|------|-----|---------------------|
| age  | 0.1380565 | 0.0507529 | 2.72 | 0.007 | 0.0385825, 0.2375304 |
| age2 | -0.0014209 | 0.0006022 | -2.36 | 0.018 | -0.0026013, -0.0002406 |
| education | 0.0898554 | 0.0272241 | 3.30 | 0.001 | 0.0364971, 0.1432137 |
| _cons | -4.714038 | 1.103265 | -4.27 | 0.000 | -6.876397, -2.551678 |

Number of obs = 433
Wald chi2(3) = 26.03
Prob > chi2 = 0.0000
Pseudo R2 = 0.0465

Log pseudolikelihood = -271.96692

Bruce Hansen (University of Wisconsin)

Bootstrapping in Stata

April 21, 2010 4 / 42
.probit mar age age2 education if bf, vce(bootstrap, reps(1000))

or

.bootstrap, reps(1000): probit mar age age2 education if bf

- The vce(bootstrap) option specifies to use the bootstrap for variance-covariance estimation (vce)
- If reps(#) is omitted, the default bootstrap replications is $R = 50$.
- The vce(bootstrap) option works with many estimation commands.
  - STATA recommends vce(bootstrap) over bootstrap as the estimation command handles clustering and model-specific details.
- bootstrap works more broadly, including non-estimation and user-written commands, or functions of coefficients.
Bootstrap Standard Error Calculation

- Computes the coefficient estimates $\hat{\beta}$ on the estimation sample
- Draws $n$ observations at random from the estimation sample
- Computes the estimates $\hat{\beta}^*$ on this simulated sample
- Repeats this $R$ times, obtaining $\hat{\beta}_b^*, b = 1, \ldots, R$

\[
\bar{\beta}^* = \frac{1}{R} \sum_{b=1}^{R} \hat{\beta}_b^*
\]

\[
se(\hat{\beta}) = \sqrt{\frac{1}{R-1} \sum_{b=1}^{R} (\hat{\beta}_b^* - \bar{\beta}^*)^2}
\]
.probit mar age age2 education if bf, vce(bootstrap, reps(1000))

Probit regression

|    | Observed Coef. | Bootstrap Std. Err. | z    | P>|z| | Normal-based [95% Conf. Interval] |
|----|----------------|---------------------|------|------|-------------------------------|
| mar|                |                     |      |      |                               |
| age| .1380565       | .0536625            | 2.57 | 0.010| .0328798  .2432331            |
| age2| -.0014209      | .0006385            | -2.23| 0.026| -.0026724 - .0001695          |
| education| .0898554      | .0279496            | 3.21 | 0.001| .0350752  .1446356            |
| _cons| -4.714038      | 1.159029            | -4.07| 0.000| -6.985693 -2.442382           |

Log likelihood = -271.96692

Number of obs = 433
Replications = 1,000
Wald chi2(3) = 25.18
Prob > chi2 = 0.0000
Pseudo R2 = 0.0465

Bootstrap standard errors.

Bruce Hansen (University of Wisconsin)  Bootstrapping in Stata  April 21, 2010 7 / 42
Function of Coefficient

In the equation

\[ \Pr(married = 1) = \Phi(\beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 edu) \]

The age at which the probability is maximized (if \( \beta_1 > 0 \) and \( \beta_2 < 0 \)) is

\[ \theta = \frac{-\beta_1}{2\beta_2} \]

This is easily estimated by

\[ \hat{\theta} = \frac{-\hat{\beta}_1}{2\hat{\beta}_2} \]

We could obtain standard errors by the Delta method. Because of the nonlinearity, bootstrap standard errors will be more reliable.
Delta Method

`.nlcom (theta: -_b[age]/(_b[age2]*2))`

. nlcom (theta: -_b[age]/(_b[age2]*2))

|       | Coef.    | Std. Err. |    z  |   P>|z| | [95% Conf. Interval] |
|-------|----------|-----------|-------|-------|----------------------|
| theta | 48.57935 | 3.873624  | 12.54 | 0.000 | 40.98719 56.17151   |

The output shows

- The estimated coefficient $\hat{\theta} = \frac{-\hat{\beta}_1}{2\hat{\beta}_2} = 49$
  - age of maximum probability of marriage
- Its asymptotic standard error
- Normal-based confidence interval
Bootstrap Standard error

`.bootstrap theta=(_b[age]/(_b[age2]*2)), reps(1000): probit mar age age2 education if bf`

\[
\hat{\theta}_b = \frac{-\hat{\beta}_{1b}}{2\hat{\beta}_{2b}}
\]

\[
se(\hat{\theta}) = \sqrt{\frac{1}{R - 1} \sum_{b=1}^{R} (\hat{\theta}_b - \bar{\theta})^2}
\]

Probit regression

| Observed Coef. | Bootstrap Coef. | z   | P>|z| | Normal-based 95% Conf. Interval |
|----------------|-----------------|-----|------|-------------------------------|
| theta          | 48.57935        | 17.42398 | 2.79 | 0.005 | (14.42898, 82.72973) |
Bootstrap Confidence Intervals

There are several methods

- Normal-based
- Percentile
- Bias-corrected (BC)
- BC$_a$ : Bias-corrected and accelerated
- Percentile-t
Choices

- STATA reports normal-based intervals in default table
  - Least desireable
- Percentile and BC intervals are easy to obtain
  - BC preferred to percentile
- The $BC_a$ is expected to perform better, but can be computationally costly in large data sets and/or non-linear estimation
- The percentile-t require more programming and requires standard errors, but can perform well
Normal Confidence interval

Stata reports a “Normal-based 95% confidence interval”:

\[ \hat{\beta} \pm 1.96 \text{se}(\hat{\beta}) \]

- It uses the bootstrap standard error
- And the asymptotic normal approximation.
- Not a good choice
Let $Q^*_\alpha$ be the $\alpha$’th quantile of $\hat{\beta}^*$

Estimated as the $\alpha$’th empirical quantile of the simulated $\hat{\beta}^*_b$

The $1 - 2\alpha$ percentile interval for $\beta$ is $[Q^*_\alpha, Q^*_1 - \alpha]$

If $\hat{\beta}$ is biased, the percentile interval will be even more biased than asymptotic interval.

Not recommended
Bias-Corrected Percentile Interval:
Define a measure of bias and its bootstrap estimator

\[ p = \Pr (\hat{\beta} < \beta) \]

\[ \hat{p} = \frac{1}{R} \sum_{b=1}^{B} \{ \hat{\beta}_b^* < \hat{\beta} \} \]

Transform \( \hat{p} \) into normal units

\[ \hat{c} = \Phi^{-1} (\hat{p}) \]

Define the shifted percentage

\[ \lambda(\alpha) = \Phi (2\hat{c} + Z_\alpha) \]

where \( Z_\alpha \) is the \( \alpha' \)th quantile of \( N(0, 1) \).
The BC percentile interval for \( \beta \) is \( [Q_{\lambda(\alpha)}^*, Q_{\lambda(1-\alpha)}^*] \)

If \( \hat{\beta} \) is unbiased then \( \hat{p} = .5, \hat{c} = 0, \lambda(\alpha) = \alpha \) and \( Q_{\lambda(\alpha)}^* = Q_\alpha^* \)

When \( \hat{\beta} \) is negatively biased then \( \hat{p} > .5, \hat{c} > 0, \lambda(\alpha) > \alpha \) and \( Q_{\lambda(\alpha)}^* > Q_\alpha^* \)

If \( \hat{\theta} \sim N(\theta - b, \sigma^2) \) is normal but biased, the BC interval will be exact.
Bias-Corrected and Accelerated Interval

$BC_a$ is similar to BC

$$\lambda(\alpha) = \Phi \left( \hat{c} + \frac{\hat{c} + Z_\alpha}{1 - a(\hat{c} + Z_\alpha)} \right)$$

When $a = 0$, then $BC$ and $BC_a$ coincide.

The optimal $a$ is called the “acceleration” because it refers to the rate of change of the standard error of $\hat{\beta}$ with respect to $\beta$, and is a scale of the skewness of $\hat{\beta}$.

$a$ is estimated in STATA by the jackknife, which requires $n$ re-estimations. If $n$ is large and estimation nonlinear this is costly.
For BC interval
```
.probit mar age age2 education if bf, vce(bootstrap, reps(1000))
estat bootstrap
```

- **estat** reports the bootstrap estimate of bias and the BC percentile interval

- A postestimation command:
  - Needs to follow estimation with bootstrap standard errors

- **estat bootstrap, all**
. estat bootstrap

Probit regression

Number of obs = 433
Replications = 1000

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>Bias</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>_cons</td>
<td>-4.7140375</td>
<td>-0.1992479</td>
<td>1.2492997</td>
<td>(-7.535929, -2.723436)</td>
</tr>
<tr>
<td>age</td>
<td>0.0898554</td>
<td>0.007389</td>
<td>0.02830316</td>
<td>(0.0320437, 0.1455764)</td>
</tr>
<tr>
<td>age2</td>
<td>-0.00142094</td>
<td>-0.0001115</td>
<td>0.00066532</td>
<td>(-0.0029528, -0.0003454)</td>
</tr>
<tr>
<td>education</td>
<td>0.13805646</td>
<td>0.0092831</td>
<td>0.05612334</td>
<td>(0.0481732, 0.2603911)</td>
</tr>
</tbody>
</table>

(BC)  bias-corrected confidence interval
. estat bootstrap, all

Probit regression

Number of obs = 433
Replications = 1000

<table>
<thead>
<tr>
<th>mar</th>
<th>Coef.</th>
<th>Bias</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>.13805646</td>
<td>.0092831</td>
<td>.05612334</td>
<td>.0280567 - .2480562 (N)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age2</td>
<td>-.00142094</td>
<td>-.0001115</td>
<td>.00066532</td>
<td>-.0027249 - .0001169 (N)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>education</td>
<td>.0898554</td>
<td>.0007389</td>
<td>.02830316</td>
<td>.0343822 - .1453286 (N)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>-4.7140375</td>
<td>-.1992479</td>
<td>1.2492997</td>
<td>-7.16262 -2.265455 (N)</td>
</tr>
</tbody>
</table>

(N) normal confidence interval
(P) percentile confidence interval
(BC) bias-corrected confidence interval
For $BC_a$ interval
\begin{verbatim}
.probit mar age age2 education if bf, vce(bootstrap, reps(1000) bca)
estat bootstrap, bca
\end{verbatim}
or
\begin{verbatim}
estat bootstrap, all
\end{verbatim}

- The **bca** option in **vce** tells STATA to calculate the acceleration $a$
- This is done by the jackknife and can be computationally costly
- The bca option in estat tells STATA to report the $BC_a$ interval instead of the BC
. estat bootstrap, bca

Probit regression

Number of obs = 433
Replications = 1000

<table>
<thead>
<tr>
<th>mar</th>
<th>Observed Coef.</th>
<th>Bias</th>
<th>Bootstrap Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>.13805646</td>
<td>.0085802</td>
<td>.05474557</td>
<td>.0327771</td>
</tr>
<tr>
<td>age2</td>
<td>-.00142094</td>
<td>-.0000992</td>
<td>.00065104</td>
<td>-.0025545</td>
</tr>
<tr>
<td>education</td>
<td>.0898554</td>
<td>.0033017</td>
<td>.02714671</td>
<td>.0373695</td>
</tr>
<tr>
<td>_cons</td>
<td>-4.7140375</td>
<td>-.2223964</td>
<td>1.1774249</td>
<td>-6.832892</td>
</tr>
</tbody>
</table>

(BCa) bias-corrected and accelerated confidence interval
```
.
.estat bootstrap, all

Probit regression

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Bias</th>
<th>Bootstrap</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td></td>
<td>Std. Err.</td>
<td></td>
</tr>
<tr>
<td>mar</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>.13805646</td>
<td>.0085802</td>
<td>.05474557</td>
<td>.0307571 .2453558 (N)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.0540528 .261395 (P)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.0467 .2478884 (BC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.0327771 .2320904 (BCa)</td>
<td></td>
</tr>
<tr>
<td>age2</td>
<td>-.00142094</td>
<td>-.0000992</td>
<td>.00065104</td>
<td>-.002697 -.0001449 (N)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-.0029675 -.0004549 (P)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-.0027947 -.0003734 (BC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.0025545 .0002154 (BCa)</td>
<td></td>
</tr>
<tr>
<td>education</td>
<td>.0898554</td>
<td>.0033017</td>
<td>.02714671</td>
<td>.0366488 .143062 (N)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.0406278 .1490357 (P)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.0374336 .1466715 (BC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.0373695 .1456622 (BCa)</td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>-4.7140375</td>
<td>-.2223964</td>
<td>1.1774249</td>
<td>-7.021748 -2.406327 (N)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-7.310817 -2.935916 (P)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-7.129984 -2.674825 (BC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-6.832892 -2.293183 (BCa)</td>
</tr>
</tbody>
</table>

(N) normal confidence interval
(P) percentile confidence interval
(BC) bias-corrected confidence interval
(BCa) bias-corrected and accelerated confidence interval
```
Function of Coefficient

- Works the same
- `.bootstrap theta=(-_b[age]/(_b[age2]*2)), reps(1000): probit mar age age2 education if bf`
- `.estat bootstrap`

```
. estat bootstrap

Probit regression

command:  probit mar age age2 education
theta:  -_b[age]/(_b[age2]*2)

<table>
<thead>
<tr>
<th></th>
<th>Observed Coef.</th>
<th>Bias</th>
<th>Bootstrap Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>theta</td>
<td>48.579351</td>
<td>.9110359</td>
<td>11.907861</td>
<td>43.86382 - 66.01958  (BC)</td>
</tr>
</tbody>
</table>

(BC) bias-corrected confidence interval
```
Alternative Syntax

`.bootstrap "probit mar age age2 education if bf" _b, reps(1000)`
or
`.bootstrap "probit mar age age2 education if bf" _b _se, reps(1000)`
contrast `.bootstrap, reps(1000): probit mar age age2 education if bf`

- Computes BC percentile intervals with one command
- Requires expression list (_b and/or _se) to specify statistics
- Works on many STATA operations
- Reports
  - bias
  - standard error
  - normal confidence interval
  - percentile interval
  - bias-corrected interval
```
.bootstrapt "probit mar age age2 education if bf" _b, reps(1000)
```

Bootstrap statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Reps</th>
<th>Observed</th>
<th>Bias</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>b_age</td>
<td>1000</td>
<td>.1380565</td>
<td>.0145857</td>
<td>.0538347</td>
<td>.0324143 .2436986</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.0603044</td>
<td>.2668444</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.0486678</td>
<td>.2495679</td>
<td></td>
</tr>
<tr>
<td>b_age2</td>
<td>1000</td>
<td>-.0014209</td>
<td>-.0001774</td>
<td>.0006448</td>
<td>-.0026863 -.0001556</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-.0029643</td>
<td>-.0005232</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-.002781</td>
<td>-.0003897</td>
<td></td>
</tr>
<tr>
<td>b_education</td>
<td>1000</td>
<td>.0898554</td>
<td>.0010376</td>
<td>.0272817</td>
<td>.0363194 1.433914</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.0394318</td>
<td>.1495175</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.0381109</td>
<td>.1486244</td>
<td></td>
</tr>
<tr>
<td>b_cons</td>
<td>1000</td>
<td>-4.714037</td>
<td>-.2999402</td>
<td>1.151227</td>
<td>-6.973137 -2.454938</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-7.56545</td>
<td>-3.009196</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-6.978446</td>
<td>-2.75727</td>
<td></td>
</tr>
</tbody>
</table>

Number of obs = 433
Replications = 1000

Bruce Hansen  
(University of Wisconsin)

Bootstrapping in Stata

April 21, 2010  25 / 42
Percentile-t Intervals

- Let $se(\hat{\beta})$ be standard error for $\hat{\beta}$
  - Best if “robust” standard error
- Let $\hat{\beta}^*, se(\hat{\beta}^*)$ be bootstrap statistics.
- Define bootstrap t-statistics

$$t^* = \frac{\hat{\beta}^* - \hat{\beta}}{se(\hat{\beta}^*)}$$

- Let $Q^*_\alpha$ be the $\alpha$'th quantile of $t^*$
- The $1 - 2\alpha$ percentile-t interval for $\beta$ is

$$[\hat{\beta} - se(\hat{\beta})Q^*_{1-\alpha}, \hat{\beta} - se(\hat{\beta})Q^*_\alpha]$$
Computation of percentile-t

- Not pre-programmed in STATA, but can be computed without programming
- Method described in Poi (STATA Journal, 2004)
- `.bootstrap "probit mar age age2 education if bf,r" _b _se, reps(1000) saving(bsdata) replace`
- In addition to earlier calculations, this stores the $R \times 1$ bootstrap statistics $\hat{\beta}^*$ and $se(\hat{\beta}^*)$ in _b and _se in the file bsdata.dta
- Be careful to specify the method for calculation of asymptotic standard errors
  - Here I use the robust option
- Next re-estimate model using original dataset using same command
- `.probit mar age age2 education if bf,r`
Load save coefficients and standard errors into memory

- `.use bsdata`

create bootstrap t-ratios

- `.gen t_age = (b_age - _b[age]) / se_age`

calculate quantiles of t-ratio (e.g. 2.5% and 97.5%)

- `_pctile t_age, p(2.5, 97.5)`

display quantiles (check results) and confidence endpoints

```
. dis r(r2)
2.4759161

. dis r(r1)
-2.0422164

. dis _b[age] - _se[age]*r(r2)
.01239647

. dis _b[age] - _se[age]*r(r1)
.24170492
```
Graphical

You can display density plots of the bootstrap distributions

- `.use bsdata`
- `.kdensity b_age`

Kernel density estimate
t-ratio

- \texttt{kdensity t\_age}

Kernel density estimate

```
Density
-4 -2 0 2 4
t\_age
kernel = epanechnikov, bandwidth = 0.2419
```

Bruce Hansen (University of Wisconsin)  Bootstrapping in Stata  April 21, 2010  30 / 42
Applying bootstrap to program output

- You can apply the bootstrap to complicated (multi-step) estimators
- You may need to write a program
- Apply the command `bootstrap` to the program
  - `.bootstrap "probit mar age" _b, reps(1000)`
  - `.bootstrap, reps(1000): probit mar age`
- Caveat: In many cases
  - STATA does not know the estimation sample
  - STATA will implement the bootstrap by drawing from all observations
  - For example, we had 207,921 total observation but only 433 in the estimation sample
  - It is best to first drop all observations which are not in the estimation sample, before running the bootstrap.
  - Otherwise, STATA will create samples of size 207,921
Joint Tests

- Wald Tests

\[ W = (\hat{\theta} - \theta_0)' \hat{V}_\theta^{-1} (\hat{\theta} - \theta_0) \]

where \( \hat{V}_\theta \) is an estimate of \( \text{var}(\hat{\theta}) \)

- Conventional Wald test uses estimate of asymptotic variance matrix
- STATA test after bootstrap estimation uses bootstrap estimate of variance matrix
- STATA does not implement correct bootstrap joint test
  - p-values based on distribution of

\[ W^* = (\hat{\theta}^* - \hat{\theta})' \hat{V}_\theta^*^{-1} (\hat{\theta}^* - \hat{\theta}) \]

- Could be done via programming
- Note that the bootstrap specifies the null values of \( \theta \) to be \( \hat{\theta} \), not \( \theta_0 \)
Quantile Regression

- Quantile Regression has its own bootstrap syntax
- Example: conditional median of hourly wage
  - function of age, education, age^2, education^2
  - same sample: black women, non-south

```
.qreg hrwage age age2 education if bf
```

```
Median regression
Number of obs = 433
Raw sum of deviations 1689.071 (about 15.264423)
Min sum of deviations 1334.419
Pseudo R2 = 0.2100

|          | Coef.   | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|----------|---------|-----------|-------|------|----------------------|
| hrwage   |         |           |       |      |                      |
| age      | 0.5230953 | 0.2507331 | 2.09  | 0.038 | 0.0302772 to 1.015913 |
| age2     | -0.0036805 | 0.0029256 | -1.26 | 0.209 | -0.0094307 to 0.0020697 |
| education| 2.330236  | 0.1704213 | 13.67 | 0.000 | 1.995271 to 2.6652   |
| _cons    | -30.9965  | 5.583125  | -5.55 | 0.000 | -41.97019 to -20.02282 |
```
Bootstrap standard errors

\[ \text{bsqreg hrwage age age2 education if bf, reps(1000)} \]

Median regression, bootstrap(1000) SEs

Raw sum of deviations 1689.071 (about 15.264423)
Min sum of deviations 1334.419

| Variable   | Coef.   | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|------------|---------|-----------|-------|-------|----------------------|
| hrwage     |         |           |       |       |                      |
| age        | .5230953| .26829    | 1.95  | 0.052 | -.0042312 1.050422   |
| age2       | -.0036805| .0033228  | -1.11 | 0.269 | -.0102116 .0028505  |
| education  | 2.330236| .2221715  | 10.49 | 0.000 | 1.893555  2.766916  |
| _cons      | -30.9965| 5.809094 | -5.34 | 0.000 | -42.41433 -19.57868 |
Functions of parameters

```
.bootstrap theta=(-_b[age]/(_b[age2]*2)), reps(1000): qreg hrwage age age2 education if bf
```

Bootstrap results

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Bootstrap</th>
<th>Normal-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>Std. Err.</td>
<td>z</td>
</tr>
<tr>
<td>theta</td>
<td>71.0625</td>
<td>1553345</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Number of obs = 433
Replications = 1,000
**BC Percentile Intervals**

*estat bootstrap* does not work with *qreg*

Instead use

```
.bstrap "qreg hrwage age age2 education if bf" _b, reps(1000)
saving (bsdata) replace
```

Bootstrap statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Reps</th>
<th>Observed</th>
<th>Bias</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>b_age</td>
<td>1000</td>
<td>.5230953</td>
<td>.0135267</td>
<td>.2661738</td>
<td>(.0007715, 1.045419) (N)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(.0582089, 1.161127) (P)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(.0525522, 1.156968) (BC)</td>
</tr>
<tr>
<td>b_age2</td>
<td>1000</td>
<td>-.0036805</td>
<td>-.000395</td>
<td>.0032746</td>
<td>(-.0101064, .0027454) (N)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(-.0119208, .001881) (P)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(-.0110044, .0024124) (BC)</td>
</tr>
<tr>
<td>b_education</td>
<td>1000</td>
<td>2.330235</td>
<td>-.0193944</td>
<td>.2206628</td>
<td>(1.89722, 2.763251) (N)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.862233, 2.706177) (P)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(1.86039, 2.705699) (BC)</td>
</tr>
<tr>
<td>b_cons</td>
<td>1000</td>
<td>-30.9965</td>
<td>.4477234</td>
<td>5.793506</td>
<td>(-42.36534, -19.62766) (N)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(-42.95274, -19.603) (P)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(-44.16151, -20.38403) (BC)</td>
</tr>
</tbody>
</table>
Plot bootstrap distributions

```
.use bsdata, replace
.kdensity b_age
```
Early literature suggested that small $R$ (e.g. $R = 50$) is sufficient.

This advice is stated in STATA manual.

Recent research (Andrews & Buchinsky) says that this is far from sufficient.

$R = 1000$ is a minimum for most calculations.

$R > 3000$ is often necessary.

I suggest using $R = 10,000$ for final calculations if possible (for submission/publication).

Andrews-Buchinsky derive methods for determining $R$.

- Might be easier to just set $R$ large and be patient.
Clustered Samples

- Bootstrap methods treat a cluster as an observation
- Resample entire clusters to create bootstrap data sets
- Example: Duflo, Dupas and Kramer (2011) investigate the impact of tracking on testscores in elementary schools in Kenya.
- 111 schools (clusters)
Linear regression, clustered by schoolid

.reg testscore tracking, cluster(schoolid)

Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
_tracking  .1469204   .0768674    1.91   0.059    -.0054128    .2992536
_cons    -.0824249   .0535249   -1.54   0.126   -.1884986    .0236488

Root MSE = .99743
R-squared = 0.0053
Prob > F = 0.0586
F(1, 110) = 3.65
Number of obs = 5,269

(Std. Err. adjusted for 111 clusters in schoolid)
Bootstrap commands

\texttt{.bootstrap, reps(1000): reg testscore tracking, cluster(schoolid)}
\texttt{.reg testscore tracking, cluster(schoolid) vce(bootstrap, reps(1000) bca)}

(Replications based on 111 clusters in schoolid)

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Bootstrap</th>
<th></th>
<th></th>
<th>Normal-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>Std. Err.</td>
<td></td>
<td></td>
<td>[95% Conf. Interval]</td>
</tr>
<tr>
<td>testscore</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tracking</td>
<td>.1469204</td>
<td>.0788651</td>
<td>1.86</td>
<td>0.062</td>
<td>-.0076523 .3014931</td>
</tr>
<tr>
<td>_cons</td>
<td>-.0824249</td>
<td>.0565709</td>
<td>-1.46</td>
<td>0.145</td>
<td>-.1933019 .0284521</td>
</tr>
</tbody>
</table>
.estat bootstrap, all

(Replications based on 111 clusters in schoolid)

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>Bias</td>
</tr>
<tr>
<td>tracking</td>
<td>0.14692041</td>
<td>-0.0016042</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>-0.08242489</td>
<td>0.0018846</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(N) normal confidence interval
(P) percentile confidence interval
(BC) bias-corrected confidence interval
(BCa) bias-corrected and accelerated confidence interval